

# Weak convergence and optimisation of the reversible jump algorithm

Philippe Gagnon,

*Université de Montréal, Canada*

Mylène Bédard

*Université de Montréal, Canada*

and Alain Desgagné

*Université du Québec à Montréal, Canada*

**Summary.** The reversible jump algorithm is a useful Markov chain Monte Carlo method introduced by [Green \(1995\)](#) that allows switches between subspaces of differing dimensionality, and therefore, model determination. Although this method is now increasingly used in key areas of human activity (e.g. finance and biology), it remains a challenge to practically and efficiently implement it. In this paper, we focus on a simple sampling context in order to obtain theoretical results that lead to optimisation of the reversible jump algorithm, and consequently, to easy implementation. The key result is the weak convergence of the sequence of stochastic processes engendered by the algorithm. This represents the main contribution of this paper as this is, to our knowledge, the first weak convergence result for the reversible jump algorithm.

**Keywords:** Markov chain Monte Carlo methods; Random walk Metropolis; Bayesian inference; Model selection; Integrated autocorrelation time; Optimal implementation

## 1. Introduction

Markov chain Monte Carlo (MCMC) methods are most commonly applied in Bayesian analysis of complex statistical models to compute estimates. They are also used to solve problems approached from a frequentist perspective, for instance in clustering (see [Kang \(2013\)](#)). In this paper, we however explain the different concepts by focusing on the primary use of these methods.

The principle of MCMC methods is to construct a Markov chain with an invariant measure that corresponds to the distribution from which we are interested in obtaining a sample (usually called the target distribution). The implementation of such samplers usually requires the specification of some functions. For instance, at each step of the Metropolis-Hastings (MH) algorithm ([Metropolis et al. \(1953\)](#) and [Hastings \(1970\)](#)), the most commonly used method, a candidate for the next state of the Markov chain is generated from a proposal distribution (which has to be specified) and accepted according to a probability function (which is provided by the method). In a Bayesian context, this means that at each step, an attempt to update the parameters is made using the proposal distribution. The specification of the required functions can be challenging for non-specialists (and even for specialists), which makes them doubt the quality of their outputs. Indeed, a poor design of these functions can lead to an inefficient algorithm, in the sense that the resulting Markov chain explores

its state space slowly, thus producing an inadequate sample (see [Peskun \(1973\)](#) and [Tierney \(1998\)](#) for a detailed explanation).

[Roberts \*et al.\* \(1997\)](#) studied the MH algorithm in the situation where the proposal distribution is a normal centered around the current state of the chain (this algorithm is a random walk Metropolis (RWM)). In this case, the specification step consists in selecting the variance of the normal distribution. This task is however not trivial as small variances lead to tiny movements of the Markov chain, while large variances induce high rejection rates of candidates. In their paper, the authors prove the existence of an optimal variance for the random walk, assuming that the algorithm is used to sample from a distribution of  $n$  independent and identically distributed (i.i.d.) random variables. They also provide a simple strategy to determine this optimal variance, which leads to a straightforward implementation of the algorithm. A lot of research has been carried out to generalise this result to more elaborate target distributions (e.g. [Roberts and Rosenthal \(2001\)](#), [Neal and Roberts \(2006\)](#), [Bédard \(2007\)](#), [Bédard \(2008\)](#), [Beskos \*et al.\* \(2009\)](#), [Bédard \*et al.\* \(2012\)](#), [Mattingly \*et al.\* \(2012\)](#) and [Beskos \*et al.\* \(2013\)](#)).

A flaw of RWM algorithms (and MH algorithms in general) is that they do not allow switches between subspaces of differing dimensionality, and therefore, model determination. This gap was corrected by [Green \(1995\)](#) with the introduction of the reversible jump algorithm. This method has a tremendous potential because of its capability to deliver information on both the “good” model and its parameters, simultaneously. For instance, [Richardson and Green \(1997\)](#) used it to estimate the number of components and the parameters of mixtures. This advantage comes with a downside: many functions have to be specified in order to do the implementation. In this paper, we study the reversible jump algorithm in the same mindset as [Roberts \*et al.\* \(1997\)](#); we aim at providing guidelines to users and open new research directions towards an automatic reversible jump algorithm.

Existing research on the reversible jump algorithm has mainly focused on ways to facilitate subspace switchings (e.g. [Brooks \*et al.\* \(2003\)](#), [Hastie \(2005\)](#), [Al-Awadhi \*et al.\* \(2004\)](#) and [Karagiannis and Andrieu \(2013\)](#)), and therefore, the exploration of the entire state space. The main drawback of the proposed approaches is the difficulty to implement them. This justifies the need for practical guidelines that will promote accessibility of the reversible jump algorithm. As a first step towards automated implementation of this method, we focus on a simple sampling context in order to obtain theoretical results. The context is defined in Section 2. The key result, which is the weak convergence of the sequence of stochastic processes engendered by the algorithm, is presented in Section 3.1. In Section 3.2, this result is used to propose an optimal design for the sampler. This is followed in Section 4.1 by a detailed procedure to implement an efficient reversible jump algorithm. In §4.1, we also discuss extensions of our results to more elaborate target distributions and formulate a conjecture. In Section 4.2, we illustrate the impact of the design of the sampler via a simulation study. Finally, the conclusion is given in Section 5. The proof of the weak convergence is substantial and can be found in Section 6. Results used in this proof that also have substantial demonstrations are presented in Section 7 to ease the reading. For the same reason, the proofs of propositions included in the text can be found in Section 8.

## 2. Sampling Context

Let

$$\pi_n(k, \mathbf{x}^k) = p(k) \prod_{i=1}^{n+k} f(x_i^k)$$

be the joint posterior distribution of  $(K^n, \mathbf{X}^{K^n})$ , where  $K^n \in \{1, \dots, \lfloor \sqrt{n} \times \log n \rfloor\}$  ( $\lfloor \cdot \rfloor$  is the floor function),  $\mathbf{X}^{K^n} := (X_1^{K^n}, \dots, X_{n+K^n}^{K^n}) \in \mathbb{R}^{n+K^n}$ ,  $n \in \{7, 8, \dots\}$ ,  $f$  is a strictly positive one-dimensional probability density function (PDF) with respect to Lebesgue measure, and  $p$  is a probability mass function (PMF) such that  $p(k) > 0$  for all  $k \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$ . The random variable  $K^n$  represents the model indicator ( $K^n = 1$  implies that model 1 is considered, for instance), and  $\mathbf{X}^{K^n}$  is the parameter vector of model  $K^n$ . Therefore,  $X_1^1, \dots, X_{n+1}^1$  are the  $n+1$  parameters of model 1,  $X_1^2, \dots, X_{n+2}^2$  are the  $n+2$  parameters of model 2, etc. To simplify the notation, we will denote  $K := K^n$  and then  $\mathbf{X}^K := \mathbf{X}^{K^n}$ . Note that the random variables  $X_1^K, \dots, X_{n+K}^K$  are conditionally i.i.d. given  $K$ , and that the random variables  $X_i^j$  and  $X_i^l$  have the same distribution, for all  $i \leq n+j$ , when  $j \leq l$ .

The objective is to obtain a representative sample from the joint posterior distribution of  $(K, \mathbf{X}^K)$  through MCMC methods in order to estimate probabilities, expectations, or any other quantity we might be interested in. MCMC users look for a simple and efficient way to attain this goal and the purpose of this paper is to provide guidelines. The following reversible jump algorithm is applied to sample from  $\pi_n$ :

- Considering that the time- $m$  state of the chain is  $(K(m), \mathbf{X}^{K(m)}(m))$ ,  $m \in \mathbb{N}$ , the type of movement that will be attempted  $J(m+1) \in \{1, 2, 3\}$  is generated from  $g$ , a PMF such that  $g(j) > 0$  for  $j \in \{1, 2, 3\}$ .
- If  $J(m+1) = 1$ , an attempt to update the parameters of the current model is made using a random walk. More precisely,  $\mathbf{Y}^{K(m)}(m+1) \sim \mathcal{N}(\mathbf{X}^{K(m)}(m), (\ell^2/(n+K(m)))\mathcal{I}_{n+K(m)})$  is generated, where  $\mathbf{Y}^{K(m)}(m+1) := (Y_1^{K(m)}(m+1), \dots, Y_{n+K(m)}^{K(m)}(m+1))$ ,  $\mathcal{I}_{n+K(m)}$  is the identity matrix of size  $n+K(m)$  and  $\ell$  is a positive constant. This candidate is accepted, i.e.  $(K(m+1), \mathbf{X}^{K(m+1)}(m+1)) = (K(m), \mathbf{Y}^{K(m)}(m+1))$ , with probability

$$1 \wedge \frac{\prod_{i=1}^{n+K(m)} f(Y_i^{K(m)}(m+1))}{\prod_{i=1}^{n+K(m)} f(X_i^{K(m)}(m))}. \quad (1)$$

- If  $J(m+1) = 2$ , an attempt to add a parameter to switch from model  $K(m)$  to model  $K(m)+1$  is made. More precisely,  $U(m+1) \sim q$  is generated and this candidate is accepted, i.e.  $(K(m+1), \mathbf{X}^{K(m+1)}(m+1)) = (K(m)+1, (\mathbf{X}^{K(m)}(m), U(m+1)))$ , with probability

$$1 \wedge \frac{f(U(m+1))p(K(m)+1)g(3)}{q(U(m+1))p(K(m))g(2)}, \quad (2)$$

where  $q$  is a strictly positive PDF.

- If  $J(m+1) = 3$ , an attempt to withdraw the last parameter to switch from model  $K(m)$  to model  $K(m)-1$  is made, i.e.  $(K(m+1), \mathbf{X}^{K(m+1)}(m+1)) = (K(m)-1, \mathbf{X}^{K(m)-}(m))$ , and this is accepted with probability

$$1 \wedge \frac{q(X_{n+K(m)}^{K(m)}(m))p(K(m)-1)g(2)}{f(X_{n+K(m)}^{K(m)}(m))p(K(m))g(3)}, \quad (3)$$

where  $\mathbf{X}^{K(m)-}(m)$  is the vector  $\mathbf{X}^{K(m)}(m)$  without the last component (more precisely  $\mathbf{X}^{K(m)-}(m) := (X_1^{K(m)}, \dots, X_{n+K(m)-1}^{K(m)})$ ).

- In case of rejection, the chain remains at the same state, i.e.  $(K(m+1), \mathbf{X}^{K(m+1)}(m+1)) = (K(m), \mathbf{X}^{K(m)}(m))$ .

Note that the resulting process  $\{(K(m), \mathbf{X}^{K(m)}), m \in \mathbb{N}\}$  is a  $\pi_n$ -irreducible and aperiodic Markov chain. In addition, it is easily shown that this Markov chain satisfies the reversibility condition with respect to  $\pi_n$  (it is nevertheless explicitly verified in [Gagnon \(2016\)](#)), and therefore, that it is ergodic, which guarantees that the Law of Large Numbers holds.

Regularity conditions imposed on the different functions are now described. They allow to obtain the theoretical results stated in Section 3.

First, we assume that the following smoothness conditions on the function  $f$  are satisfied:  $f \in \mathcal{C}^2(\mathbb{R})$  (the space of real-valued functions on  $\mathbb{R}$  with continuous second derivative),  $(\log f(x))'$  is Lipschitz continuous and  $\mathbb{E}[(\log f(X))'^4] < \infty$ , where the expectation is computed with respect to  $f$ . This last condition can be replaced by  $\mathbb{E}[(f''(X)/f(X))^2] < \infty$ , which is slightly stronger. We also assume that there exists a constant  $A^* \geq 1$  such that

$$0 < \frac{f}{q} \leq A^* \text{ and therefore } \frac{1}{A^*} \leq \frac{q}{f} < \infty.$$

This condition corresponds to that required for the rejection sampling method. It ensures that the tails of  $q$  are at least as heavy as those of  $f$ , and thus, that  $q$  induces a good exploration of the state space. A small value for the constant  $A^*$  means that  $q$  is similar to  $f$ , and therefore, that it is a good choice of proposal distribution. Note that, when we can directly sample from  $f$ , we can set  $q = f$ .

The distribution  $p$  also fulfills some conditions. We assume that the mode of this distribution is in the middle of the set  $\{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$  and that this PMF is symmetric with respect to this mode. Two distinct cases thus have to be considered: when  $\lfloor \sqrt{n} \log n \rfloor$  is even or odd. When  $\lfloor \sqrt{n} \log n \rfloor$  is odd, the mode is  $(\lfloor \sqrt{n} \log n \rfloor + 1)/2$  and we assume that

$$\begin{aligned} p(k+1) &= a_{k,n} p(k), k \in \{(\lfloor \sqrt{n} \log n \rfloor + 1)/2, \dots, \lfloor \sqrt{n} \log n \rfloor - 1\}, \\ p(k-1) &= a_{k-1,n} p(k), k \in \{2, \dots, (\lfloor \sqrt{n} \log n \rfloor + 1)/2\}, \end{aligned}$$

where  $a_{k,n} := (1 - b_{k,n}/\sqrt{n})$  with

$$b_{k,n} := \left| \frac{k - \lfloor \sqrt{n} \log n \rfloor / 2}{\sqrt{n}} \right|.$$

Note that  $a_{k,n}$  decreases with the distance between  $k$  and the mode. This distribution is symmetric with respect to  $(\lfloor \sqrt{n} \log n \rfloor + 1)/2$  and is such that

$$p\left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} + k\right) = p\left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} - k\right) = p\left(\frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2}\right) \prod_{i=1}^k \left(1 - \frac{i-1/2}{n}\right),$$

where  $k \in \{1, \dots, (\lfloor \sqrt{n} \log n \rfloor - 1)/2\}$ .

When  $\lfloor \sqrt{n} \log n \rfloor$  is even, the distribution  $p$  is bimodal with modes at  $\lfloor \sqrt{n} \log n \rfloor / 2$  and  $\lfloor \sqrt{n} \log n \rfloor / 2 + 1$ . Using the same definitions as above for  $a_{k,n}$  and  $b_{k,n}$ , we assume that

$$\begin{aligned} p(k+1) &= a_{k,n} p(k), k \in \{\lfloor \sqrt{n} \log n \rfloor / 2 + 1, \dots, \lfloor \sqrt{n} \log n \rfloor - 1\}, \\ p(k-1) &= a_{k-1,n} p(k), k \in \{2, \dots, \lfloor \sqrt{n} \log n \rfloor / 2\}, \end{aligned}$$

which implies that

$$p\left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2} + 1 + k\right) = p\left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2} - k\right) = p\left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2}\right) \prod_{i=1}^k \left(1 - \frac{i}{n}\right), \quad (4)$$

where  $k \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor / 2 - 1\}$ .

The assumptions on  $p$  imply that its mode is in the middle of its domain, with probabilities that decrease with the distance to this mode at an exponential rate which is bounded below by  $1/2$  (the ratios  $p(k+1)/p(k)$  and  $p(k-1)/p(k)$  are essentially bounded below by  $1/2$ ). To fix ideas, consider the case where  $\lfloor \sqrt{n} \log n \rfloor = 5$  ( $n = 7$ ). The shape of  $p$  is such that the “best” model has  $n + (\lfloor \sqrt{n} \log n \rfloor + 1)/2 = 10$  parameters, while the model with an additional parameter is less appropriate (so is the model with one less parameter), in the sense that  $p(4)/p(3) = p(2)/p(3) = 0.93$ . The more parameters we add (or withdraw), the less appropriate the models are. This structure becomes natural when reversible jump users rank the models by number of parameters and believe the posterior distribution of models reflects the existence of a balance between a good fit (which involves a lot of parameters), and simplicity and stability of models, a principle aligned with Occam’s razor. In addition, the shape of the PMF  $p$  leads to an efficient exploration of the entire state space, because if the ratios  $p(k+1)/p(k)$  and  $p(k-1)/p(k)$  were close to 0 for some values of  $k$ , there would be more rejected attempts of switches from model  $k$  to model  $k+1$  or  $k-1$  (see (2) and (3)).

The hypothesised mathematical structure of the PMF  $p$  allows to obtain theoretical results, as (see Proposition 1 in Section 3.1 for the formal statement)

$$\mathbb{P}\left(\frac{K - \frac{\lfloor \sqrt{n} \log n \rfloor}{2}}{\sqrt{n}} \leq x\right) \rightarrow \Phi(x), \forall x \in \mathbb{R}, \text{ as } n \rightarrow \infty, \quad (5)$$

where  $K \sim p$  and  $\Phi$  is the cumulative distribution function of the standard normal. It can be proved that, for all values of  $k$  such that  $b_{k,n}$  is well-defined,  $b_{k,n} \leq \log(n)/2$  and, therefore, that  $b_{k,n}/\sqrt{n} \rightarrow 0$  as  $n \rightarrow \infty$ , which implies that  $a_{k,n} \rightarrow 1$  as  $n \rightarrow \infty$ . Moreover, for all  $n$  and for all values of  $k$  such that  $b_{k,n}$  is well-defined,  $0 < b_{k,n}/\sqrt{n} \leq 1/2$ , which implies that  $1/2 \leq a_{k,n} < 1$ . Thus, when  $k \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor - 1\}$ , we have that  $1/2 \leq p(k+1)/p(k) \leq 2$  and  $p(k+1)/p(k) \rightarrow 1$  as  $n \rightarrow \infty$ , because this ratio is essentially equal to  $a_{k,n}$  or  $a_{k,n}^{-1}$ . To summarise, the assumptions on  $p$  and the standardisation of  $K$  in (5) are such that the resulting random variable is continuous and takes values on the real line, in the limit, which makes this convergence in distribution possible. Note that the assumptions on  $p$  indeed imply that  $p(k) > 0$  for all  $k \in \{1, \dots, \lfloor \sqrt{n} \log n \rfloor\}$ .

Finally, we define the function  $g$  as follows:

$$g(j) := \begin{cases} \tau & \text{if } j = 1, \\ (1 - \tau)A/(A + 1) & \text{if } j = 2, \\ (1 - \tau)/(A + 1) & \text{if } j = 3, \end{cases} \quad (6)$$

where  $0 < \tau < 1$  is a constant and  $A := 2A^*$ . Considering this definition, the acceptance probability arising from the inclusion of an extra parameter (see (2)) becomes the minimum between 1 and  $f(U)/q(U) \times 1/A \times p(K+1)/p(K)$ . By assumption,  $2f/q \leq A$  and  $p(K+1)/p(K) \leq 2$ ; therefore, this acceptance probability is simply  $f(U)/q(U) \times 1/A \times p(K+1)/p(K)$ , which is easier to handle mathematically than the minimum function expressed in (2). Furthermore, the acceptance probability arising from the withdrawal of the last parameter (see (3)) becomes the minimum between 1

and  $q(X_{n+K}^K)/f(X_{n+K}^K) \times A \times p(K-1)/p(K) \geq 1$ , which means that this type of movement is automatically accepted (whenever it is possible to withdraw a parameter, i.e. when  $K > 1$ ).

### 3. Towards Optimal Implementation of the Reversible Jump

In order to implement the reversible jump algorithm described in Section 2, we have to specify the PDF  $q$  and values for the constants  $A$ ,  $\tau$  and  $\ell$ . In Section 3.1, we present weak convergence results that are used in Section 3.2 to find asymptotically optimal values for  $\tau$  and  $\ell$ . In Section 4.1, we provide guidelines to suitably design  $q$ , which implicitly allows to determine the constant  $A$ .

#### 3.1. Weak Convergence Results

In order to study the asymptotic behaviour of the algorithm, we consider the following rescaled stochastic process:

$$\mathbf{Z}^n(t) := \left( \frac{K(\lfloor nt \rfloor) - \lfloor \sqrt{n} \log n \rfloor / 2}{\sqrt{n}}, \mathbf{X}^{K(\lfloor nt \rfloor)}(\lfloor nt \rfloor) \right), \quad (7)$$

where  $t \geq 0$ . The continuous-time stochastic process  $\{\mathbf{Z}^n(t), t \geq 0\}$  is a sped up and modified version of  $\{(K(m), \mathbf{X}^K(m)), m \in \mathbb{N}\}$  (see Section 2 for the definition of this process) that admits jumps. In a given iteration, the average distance travelled by the parameters

$$\mathbf{X}^{K(\lfloor nt \rfloor)}(\lfloor nt \rfloor) := (X_1^{K(\lfloor nt \rfloor)}(\lfloor nt \rfloor), \dots, X_{n+K(\lfloor nt \rfloor)}^{K(\lfloor nt \rfloor)}(\lfloor nt \rfloor))$$

decreases with  $n$  because the variance of the random walk is proportional to  $1/(n + K(\lfloor nt \rfloor))$ . In addition, the distance travelled by  $\{K(\lfloor nt \rfloor)/\sqrt{n}, t \geq 0\}$ , each time it moves, is  $1/\sqrt{n}$ . The decreasing size of the jumps, combined with the acceleration of  $\{\mathbf{Z}^n(t), t \geq 0\}$ , result in a continuous and non-trivial limiting process. As explained in Section 2, we subtract  $\lfloor \sqrt{n} \log n \rfloor / (2\sqrt{n})$  from  $\{K(\lfloor nt \rfloor)/\sqrt{n}, t \geq 0\}$  in order to obtain a limiting process with components that take values on the real line. The asymptotic behaviour of  $\{Z_1^n(t), t \geq 0\}$  is described precisely in Proposition 1.

**Proposition 1.** *Consider the context described in Section 2, the stochastic process  $\{\mathbf{Z}^n(t), t \geq 0\}$  defined in (7), and assume that  $\mathbf{Z}^n(0) \sim \pi_n$ . Then, as  $n \rightarrow \infty$ ,  $Z_1^n(t)$  converges in distribution towards a standard normal random variable, for all  $t \geq 0$ .*

*Proof.* See Section 8. ■

The main result is now stated.

**Theorem 1.** *Consider the context described in Section 2, the stochastic process  $\{\mathbf{Z}^n(t), t \geq 0\}$  defined in (7), and assume that  $\mathbf{Z}^n(0) \sim \pi_n$ . Then, as  $n \rightarrow \infty$ , the first two components of  $\{\mathbf{Z}^n(t), t \geq 0\}$  converge weakly towards a bidimensional Langevin diffusion, i.e.*

$$\{\mathbf{Z}_{1,2}^n(t), t \geq 0\} := \{(Z_1^n(t), Z_2^n(t)), t \geq 0\} \Rightarrow \{\mathbf{Z}(t), t \geq 0\} \text{ as } n \rightarrow \infty,$$

where the process  $\{\mathbf{Z}(t), t \geq 0\}$  is comprised of two independent components such that  $Z_1(0) \sim \mathcal{N}(0, 1)$ ,  $Z_2(0) \sim f$ ,

$$\begin{aligned} dZ_1(t) &= \sqrt{2(1-\tau)/(A+1)} dB_1(t) - (1-\tau)/(A+1) \times Z_1(t) dt, \\ dZ_2(t) &= \sqrt{2\tau\ell^2\Phi(-\ell\sqrt{\Upsilon}/2)} dB_2(t) + \tau\ell^2\Phi(-\ell\sqrt{\Upsilon}/2)(\log f(Z_2(t)))' dt, \end{aligned}$$

with  $\{B_1(t), t \geq 0\}, \{B_2(t), t \geq 0\}$  being two independent Wiener processes and

$$\Upsilon := \mathbb{E} [(\log f(Z_2(0)))'^2].$$

*Proof.* See Section 6. ■

The notation “ $\Rightarrow$ ” represents weak convergence (or convergence in distribution) of processes in the Skorokhod topology (for more details about this type of convergence, see Section 3 of [Ethier and Kurtz \(1986\)](#)).

### 3.2. Optimisation

The sample paths of  $\{\mathbf{Z}(t), t \geq 0\}$  depend on  $\tau, A, \ell, \Upsilon$  and  $f$ . In this section, we optimise theoretically the state space exploration of  $\{\mathbf{Z}(t), t \geq 0\}$  with respect to  $\ell$  and  $\tau$ . As a result, in the situation where the dimension of the models is large enough, i.e. for large enough  $n$ , reversible jump users will know how to choose the values of  $\ell$  and  $\tau$  in order to obtain the optimal algorithm. Indeed, optimising the asymptotic state space exploration of  $\{(K(\lfloor nt \rfloor), X_1^{K(\lfloor nt \rfloor)}(\lfloor nt \rfloor)), t \geq 0\}$  is sufficient to optimise the asymptotic state space exploration of  $\{(K(\lfloor nt \rfloor), \mathbf{X}^{K(\lfloor nt \rfloor)}(\lfloor nt \rfloor)), t \geq 0\}$ . This is due to the fact that, in addition to optimising the exploration of models, we also optimise the state space exploration of the first parameter, and all parameters of a model share a similar behaviour.

During the theoretical optimisation, the constant  $A$  is considered to be fixed because its value cannot be arbitrarily chosen. Indeed, it is tied to the ratio  $f/q \leq A^* = A/2$ . In addition, everything suggests that selecting a small value for this constant is desirable. The constant  $\Upsilon$  and the function  $f$  are obviously fixed. Note that the PDF  $q$  only has an impact on the sample paths of  $\{\mathbf{Z}(t), t \geq 0\}$  through the constant  $A$ .

We first optimise the algorithm with respect to  $\ell$ . The processes  $\{Z_1(t), t \geq 0\}$  and  $\{Z_2(t), t \geq 0\}$  have speed measures given by  $2(1 - \tau)/(A + 1)$  and  $2\tau\ell^2\Phi(-\ell\sqrt{\Upsilon}/2)$ , respectively. A speed measure of  $2\tau\ell^2\Phi(-\ell\sqrt{\Upsilon}/2)$ , considering  $\{Z_2(t), t \geq 0\}$  as an example, means that  $Z_2(t) = V(2 \times \tau\ell^2\Phi(-\ell\sqrt{\Upsilon}/2) \times t)$ , where  $\{V(t), t \geq 0\}$  is the Langevin diffusion with speed measure unity, i.e.

$$dV(t) = dB_2(t) + (\log f(Z_2(t)))' / 2 \times dt.$$

Viewed as a function of  $\tau$  and  $\ell$ , the process  $\{Z_2(t), t \geq 0\}$  that optimally explores its state space is thus the one with the largest speed. We can therefore optimise the algorithm with respect to  $\ell$  by maximising the speed of  $\{Z_2(t), t \geq 0\}$  with respect to this variable, because the value of  $\ell$  does not have an impact on the sample paths of  $\{Z_1(t), t \geq 0\}$ . The function  $2\tau\ell^2\Phi(-\ell\sqrt{\Upsilon}/2)$  is maximised with respect to  $\ell$  by  $\ell = 2.38/\sqrt{\Upsilon}$ , as stated in Corollary 1.2 of [Roberts et al. \(1997\)](#). We therefore obtain the same optimal value as these authors. This conclusion does not come as a surprise, since updating the parameters in our reversible jump algorithm (see Section 2) corresponds to the usual RWM step studied by these authors. Furthermore, the conditional distribution of the parameters given a model  $K = k$  is essentially the same as their target distribution.

The optimisation with respect to  $\ell$  tells us that the most efficient way to update the parameters is to set  $\ell = 2.38/\sqrt{\Upsilon}$ . Therefore, the optimal variance for the random walk is  $(2.38^2/(\Upsilon(n + K(m))))\mathcal{I}_{n+K(m)}$ . It could seem necessary to know  $\Upsilon = \mathbb{E}[(\log f(Z_2(0)))'^2]$  in order to use this optimal scaling result. Fortunately, the practical 0.234 rule provided by [Roberts et al. \(1997\)](#) can be employed, as stated in Corollary 1. This corollary is an adapted version of Corollary 1.2 of [Roberts et al. \(1997\)](#). Its proof is similar to the one given by these authors, and is thus omitted (it can nevertheless be found in [Gagnon \(2016\)](#)).

**Corollary 1.** *In the context described in Section 2, assume that  $(K(0), \mathbf{X}^K(0)) \sim \pi_n$ . Then, for all  $m \in \mathbb{N}$ ,*

$$\mathbb{E} \left[ 1 \wedge \prod_{i=1}^{n+K(m)} \frac{f(Y_i^{K(m)}(m+1))}{f(X_i^{K(m)}(m))} \right] \rightarrow 2\Phi(-\ell\sqrt{\Upsilon}/2) \text{ as } n \rightarrow \infty.$$

*In addition, setting  $\ell = 2.38/\sqrt{\Upsilon}$  is equivalent to having  $2\Phi(-\ell\sqrt{\Upsilon}/2) = 2\Phi(-2.38/2) = 0.234$ .*

Therefore, in order to reach optimal efficiency with respect to  $\ell$ , reversible jump users can monitor the acceptance rate of candidates  $\mathbf{Y}^{K(m)}(m+1)$ , where  $\mathbf{Y}^{K(m)}(m+1) \sim \mathcal{N}(\mathbf{X}^{K(m)}(m), (\ell^2/(n+K(m)))\mathcal{I}_{n+K(m)})$ , and tune the value of  $\ell$  such that this rate is approximately 0.234. Note that this rate must be computed by considering only iterations in which there have been an attempt of updating the parameters, i.e. iterations belonging to the set  $\{m : J(m) = 1\}$ .

Note that the speed measures of  $\{Z_1(t), t \geq 0\}$  and  $\{Z_2(t), t \geq 0\}$  theoretically confirm that users should select the smallest value for  $A$  satisfying  $f/q \leq A^* = A/2$ . Indeed, the speed measure of  $\{Z_1(t), t \geq 0\}$ , which is given by  $2(1-\tau)/(A+1)$ , is maximised when  $A$  is small, and the one of  $\{Z_2(t), t \geq 0\}$  does not depend on  $A$ .

We now optimise the algorithm with respect to  $\tau$ . We need a measure that takes into account the fact that an increase in the value of  $\tau$  results in an increase in the speed of  $\{Z_2(t), t \geq 0\}$ , but also in a decrease in the speed of  $\{Z_1(t), t \geq 0\}$ , and vice versa. Intuitively, when the value of  $\tau$  is increased, more updates of the parameters (and therefore less model switchings) are proposed. It would seem natural to consider the total speed of these two processes to optimise the algorithm with respect to  $\tau$ . The total speed is given by (using the optimal value for  $\ell$ )

$$2[\tau(2.38^2/\Upsilon)\Phi(-2.38/2) + (1-\tau)/(A+1)].$$

However, it is not a suitable measure because if, for instance  $(2.38^2/\Upsilon) \times \Phi(-2.38/2) > 1/(A+1)$ , it would be proposed to choose the value of  $\tau$  as close as possible to 1. In such a situation, there would be very few model switchings, which would result in a slow exploration of the entire state space.

We thus need a measure that penalises such a behaviour. This is achieved using integrated linear combinations of the autocorrelation functions (ACFs) of  $\{Z_1(t), t \geq 0\}$  and  $\{Z_2(t), t \geq 0\}$ . Indeed, if for instance we set  $\tau$  close to 1,  $\{Z_2(t), t \geq 0\}$  would be a “fast” process with an ACF that decreases rapidly towards 0, while  $\{Z_1(t), t \geq 0\}$  would be a “slow” process with an almost constant ACF around the value 1 (which is not desirable). Therefore, the sum of these two functions would decrease rapidly towards 1, thereafter remaining almost constant around this value. There should then exist a value of  $\tau$  between 0 and 1 that induces two relatively “fast” processes, with a sum of ACFs that decreases relatively rapidly towards 0. We thus consider the integral of the sum of the ACFs of  $\{Z_1(t), t \geq 0\}$  and  $\{Z_2(t), t \geq 0\}$  to optimise the algorithm with respect to  $\tau$ :

$$\int_0^\infty \{\text{corr}[Z_1(t), Z_1(t+s)] + \text{corr}[Z_2(t), Z_2(t+s)]\} ds, \quad t \geq 0. \quad (8)$$

This measure is inspired from the effective sample size (see Section 12.3.5 of [Robert and Casella \(2004\)](#)) and can be viewed as the sum of the (infinitesimally) integrated autocorrelation times of  $\{Z_1(t), t \geq 0\}$  and  $\{Z_2(t), t \geq 0\}$ . It therefore represents a measure of the total “inefficiency” of these processes and the optimal value of  $\tau$  is the one that minimises it.

We need to compute the ACFs of  $\{Z_1(t), t \geq 0\}$  and  $\{Z_2(t), t \geq 0\}$  in order to optimise the algorithm with respect to  $\tau$ . The process  $\{Z_1(t), t \geq 0\}$  satisfies the conditions of Theorem 2.1



stated by Bibby *et al.* (2005), implying that

$$\text{corr}[Z_1(t), Z_1(t+s)] = \exp\{-(1-\tau)s/(A+1)\}, \quad s, t \geq 0.$$

The behaviour of  $\{Z_2(t), t \geq 0\}$  depends on  $f$  and consequently its ACF cannot be computed in all generality. A particular situation is now studied in order to obtain general information about the optimal value of  $\tau$ . It is natural to consider the case where  $f = \mathcal{N}(\mu, \sigma^2)$ ,  $\mu \in \mathbb{R}, \sigma > 0$ , since it implies that the stationary distributions of  $\{Z_1(t), t \geq 0\}$  and  $\{Z_2(t), t \geq 0\}$  are respectively a standard normal and a  $\mathcal{N}(\mu, \sigma^2)$ . Therefore, it represents a situation where all the components of  $\{\mathbf{Z}^n(t), t \geq 0\}$  have a similar behaviour when  $n$  is large enough. Using the optimal value for  $\ell$ , we have that

$$dZ_2(t) = \sqrt{2\tau(2.38^2\sigma^2)\Phi(-2.38/2)}dB_2(t) - \tau(2.38^2\sigma^2)\Phi(-2.38/2) \times (Z_2(t) - \mu)/\sigma^2 dt.$$

This process also satisfies the conditions of Theorem 2.1 stated by Bibby *et al.* (2005), implying that

$$\text{corr}[Z_2(t), Z_2(t+s)] = \exp\{-\tau 2.38^2 \Phi(-2.38/2) s\}, \quad s, t \geq 0.$$

Thus, when  $f = \mathcal{N}(\mu, \sigma^2)$  and  $\ell$  is set to its optimal value, the optimal value for  $\tau$  is

$$\text{argmin}_{0 < \tau < 1} \frac{2.38^2 A \Phi(-2.38/2) \tau + \tau(2.38^2 \Phi(-2.38/2) - 1) + 1}{2.38^2(1-\tau)\tau\Phi(-2.38/2)}. \quad (9)$$

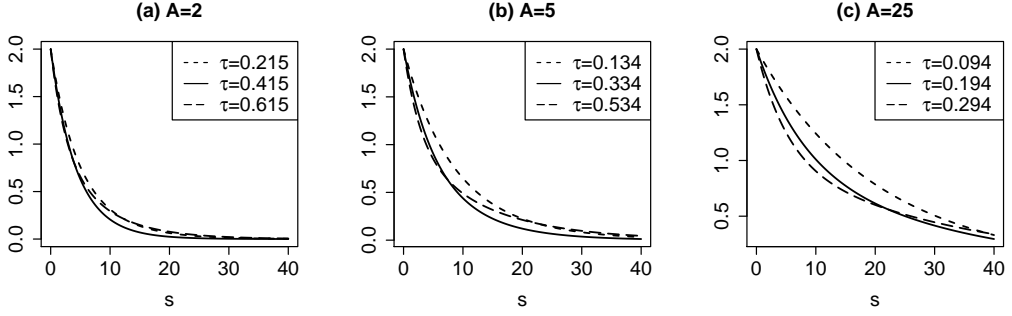
The constant  $A$  clearly has an impact on the optimal value of the constant  $\tau$ . In fact, the optimal value decreases as the value of  $A$  increases, as depicted in Figures 1 and 2. Note that the sum of the ACFs arising from the optimal value of the constant  $\tau$  represents the curve which decreases most rapidly towards 0, in the sense that the area under the curve is minimised. It indicates that the underlying process  $\{\mathbf{Z}(t), t \geq 0\}$  optimally explores its state space. When  $A = 2$ , the optimal value for  $\tau$  is 0.415. This situation corresponds to  $f = q$ , and therefore to the best choice of distribution  $q$ . When  $A = 5$ , the optimal value for  $\tau$  is 0.334, and when  $A = 25$ , it is 0.194. The constant  $A$  therefore has an indirect impact on the sample paths of  $\{Z_2(t), t \geq 0\}$  when  $\tau$  is set to its optimal value. Indeed, the speed measure of this process, which is given by  $2\tau\ell^2\Phi(-\ell\sqrt{\Upsilon}/2)$ , decreases as  $A$  increases if  $\tau$  is set to its optimal value. Again, selecting the smallest admissible value for  $A$  is desirable.

In the situation where we cannot directly sample from  $f$ , the constant  $A$  represents in some way the level of precision in the design of the distribution  $q$ . For illustrative purpose, assume that  $f = \mathcal{N}(\mu_1, \sigma_1^2)$  and that a user considers the proposal distribution  $q = \mathcal{N}(\mu_2, \sigma_2^2)$ . If he believes he might have overestimated the variability by a factor of at most 1.5 ( $\sigma_2/\sigma_1 \leq 1.5$ ) and the location by at most 1 ( $0 \leq \mu_2 - \mu_1 \leq 1$ ), then, provided that  $\sigma_2^2 \geq 1 + \sigma_1^2$ , he should set the constant  $A$  to 5 (see Figure 1 (b) for the impact on the sum of the ACFs). Indeed,

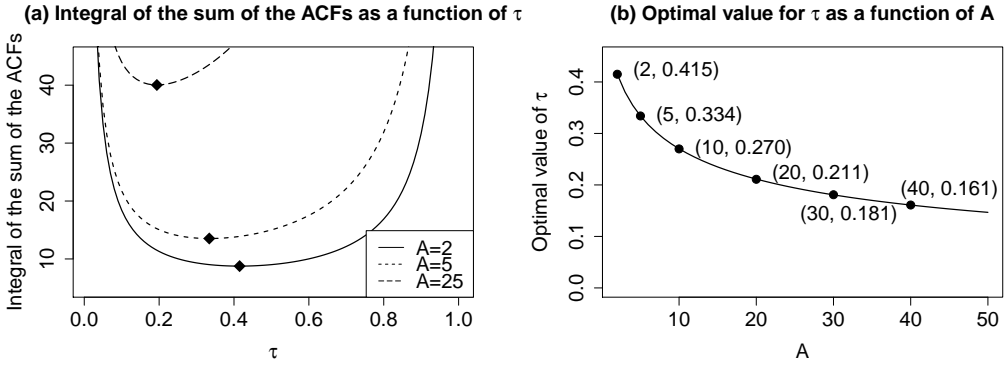
$$\frac{f(x)}{q(x)} \leq \max_x \frac{f(x)}{q(x)} = \frac{\sigma_2}{\sigma_1} \exp\left\{\frac{1}{2} \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2 - \sigma_1^2}\right\} \leq 2.47 \leq A^* =: A/2.$$

Note that this upper bound is valid only if  $\sigma_2 > \sigma_1$ ; otherwise, the ratio  $f/q$  is unbounded. It means that users should always be conservative towards variability, in particular when they lack information about the location.

Suppose that a user sets  $\ell$  to its optimal value and, say,  $A = 5$  (while, theoretically,  $A \geq 5$ , which means that the smallest admissible value is selected). The optimisation with respect to  $\tau$  tells us that



**Fig. 1.** Sum of the ACFs as a function of  $s$  for different values of  $\tau$  and  $A = 2, 5, 25$ , when  $f = \mathcal{N}(\mu, \sigma^2)$  and  $\ell$  is set to its optimal value (for each graph, the solid line represents the function arising from the optimal value of the constant  $\tau$ )



**Fig. 2.** (a) Integral of the sum of the ACFs as a function of  $\tau$ , for  $A = 2, 5, 25$  (the diamonds represent the optimal value for  $\tau$ ), (b) Optimal value for  $\tau$  as a function of  $A$ ; for both graphs,  $f = \mathcal{N}(\mu, \sigma^2)$  and  $\ell$  is set to its optimal value

the optimal distribution  $g^*$  is given by

$$g^*(j) = \begin{cases} \tau = 0.33 & \text{if } j = 1, \\ (1 - \tau)A/(A + 1) = (1 - 0.33)5/(5 + 1) = 0.56 & \text{if } j = 2, \\ (1 - \tau)/(A + 1) = (1 - 0.33)/(5 + 1) = 0.11 & \text{if } j = 3, \end{cases}$$

assuming that  $f = \mathcal{N}(\mu, \sigma^2)$ . Recall that the random variable  $J(m + 1), m \in \mathbb{N}$ , is distributed according to  $g$  and this random variable indicates which movement type is attempted at iteration  $m + 1$ : update of the parameters ( $J(m + 1) = 1$ ), inclusion of an extra parameter ( $J(m + 1) = 2$ ) or withdrawal of the last parameter ( $J(m + 1) = 3$ ).

Upon examination of the optimal distribution  $g^*$  given above, the probabilities  $g^*(2)$  and  $g^*(3)$  might appear unbalanced. We should however focus on the probabilities of the actual movements of  $\{K(m), m \in \mathbb{N}\}$ . Intuitively, it seems effective that movements of type  $K(m) \mapsto K(m) + 1$  (inclusion of an extra parameter) be as frequent as those of type  $K(m) \mapsto K(m) - 1$  (withdrawal

of the last parameter). Given that there is an attempt to include an extra parameter (i.e. given that  $J(m+1) = 2$ ), Proposition 2 indicates that the average acceptance probability of the candidate  $U(m+1)$ , where  $U(m+1) \sim q$ , converges towards  $1/A$  as  $n \rightarrow \infty$ .

**Proposition 2.** *Consider the context described in Section 2 and the function  $g$  defined in (6). If we assume that  $(K(0), \mathbf{X}^K(0)) \sim \pi_n$ , then for all  $m \in \mathbb{N}$ ,*

$$\mathbb{E} \left[ 1 \wedge \frac{f(U(m+1))p(K(m)+1)g(3)}{q(U(m+1))p(K(m))g(2)} \right] \rightarrow \frac{1}{A} \text{ as } n \rightarrow \infty.$$

*Proof.* See Section 8. ■

It means that the average probability of a movement of type  $K(m) \mapsto K(m) + 1$  is asymptotically  $(1 - \tau)A/(A + 1) \times 1/A = (1 - \tau)/(A + 1)$ . This is equal to the average probability of a movement of type  $K(m) \mapsto K(m) - 1$ . Indeed, the probability to withdraw the last parameter in a given iteration is  $(1 - \tau)/(A + 1)$  for all  $n$  (this movement is automatically accepted, as explained in Section 2).

## 4. Practical Considerations

### 4.1. Optimal Implementation and Generalisation

When users are able to state that  $\pi_n(k, \mathbf{x}^k) = p(k) \prod_{i=1}^{n+k} f(x_i^k)$ , with  $f = \mathcal{N}(\mu, \sigma^2)$  and  $p$  defined as in Section 2, they can directly construct the optimal algorithm setting  $q = f$  (therefore  $A = 2$ ),  $\tau = 0.415$  (the optimal value for  $\tau$  given by (9) or in Figure 2 (b)), and  $\ell = 2.38\sigma$  (the optimal value for  $\ell$ ). This situation is however unlikely to occur. Our recommendation for optimising the algorithm is the following: perform some trial runs to tune the value of  $\ell$  (according to Corollary 1), and to obtain general information about  $f$  (which is useful when we cannot directly sample from  $f$ ). The information gathered about  $f$  enables to improve the design of the proposal distribution  $q$ , and thus to reduce the value of the upper bound of  $f/q$ , given by  $A^* =: A/2$ , which in turn leads to a new optimal value for  $\tau$ . The optimal value for  $\tau$  given by (9) (or in Figure 2 (b)) is theoretically valid when  $f = \mathcal{N}(\mu, \sigma^2)$ , but it should be suitable if  $f$  has a similar shape. If at the beginning of the process users lack information about  $f$ , they should start with conservative values for  $A$  and  $\ell$ .

The optimal scaling result of Roberts *et al.* (1997) is known to be relatively robust, in the sense that it holds under weaker assumptions (see, e.g., Roberts and Rosenthal (2001) and Bédard (2007)). We believe that the results presented in this paper are also robust and we conjecture that they are valid when

$$\pi_n(k, \mathbf{x}^k) = p(k) \prod_{i=1}^{n+k} f_i(x_i^k),$$

where  $f_i(x_i^k) := (1/\sigma_i)f((x_i^k - \mu_i)/\sigma_i)$ ,  $\mu_i \in \mathbb{R}$  and  $\sigma_i > 0$  are constants, and  $f$  and  $p$  satisfy the assumptions described in Section 2. The sampler described in Section 2 would be applied, the only difference being that a proposal distribution  $q_{K+1}$  would be used to add a parameter to switch from model  $K$  to model  $K + 1$ , in order to accommodate for the different functions  $f_i$ . We would assume that  $f_i/q_i \leq A_i^* \leq A_n^*$  for all  $i \in \{n+1, \dots, n + \lfloor \sqrt{n} \log n \rfloor\}$  with  $A_n^* := \max_i A_i^*$ , and  $A_n^* \leq A^*$  for all  $n$ . Considering instead this constant  $A^*$ , we would perform the procedure given above for optimising the algorithm. The results presented in this paper would therefore be useful when, given a model, the parameters are independent but not identically distributed. In particular, they would be useful to select a subset of the principal components when the principal component regression is used to model the data. Note that the generalisation of these results to these contexts is not trivial.

#### 4.2. Simulation Study

Samples produced by the reversible jump algorithm provide information about the joint posterior distribution of  $K$ , the model indicator, and  $\mathbf{X}^K$ , the parameters of model  $K$ . As a result, users can choose a model (usually the one with the highest frequency in the sample) and estimate its parameters (using sample means and intervals for instance). Ideally, the chosen model, along with the parameter estimates, would be the same as if the “true” posterior distribution had been used. In Section 3, we have explained how to implement an efficient reversible jump algorithm using optimisation results that have been derived from Theorem 1. In this section, we illustrate the impact of the design of the sampler on the estimation of the target distribution via a simulation study.

Implementing the reversible jump algorithm described in Section 2 comes down to specifying the PDF  $q$  and the values of the constants  $A, \tau$  and  $\ell$ . In Section 3, it has been shown that the constants  $\tau$  and  $A$  have an impact on the estimation of the whole joint posterior distribution of  $(K, \mathbf{X}^K)$ . The constant  $\ell$  has an impact on the  $\mathbf{X}^K$  part only and this has been thoroughly studied by Roberts and Rosenthal (2001). In Section 3, it has also been explained that, asymptotically, the PDF  $q$  only has an impact through the constant  $A$ . In this section, we therefore focus on showing the impact of the constants  $\tau$  and  $A$  on the estimation of the target distribution. More precisely, considering that  $f = \mathcal{N}(\mu, \sigma^2)$ ,  $\mu \in \mathbb{R}, \sigma > 0$ , we set  $\ell = 2.38\sigma$  (its optimal value),  $q = f$ , and we evaluate the performance of the algorithm for every  $\tau \in (0, 1)$ , in the cases where  $A = 2, 5, 25$ . For a given  $A$ , the optimal value for  $\tau$  can thus be determined using (9) (or by looking at Figure 2 (b)). Studying specific situations like this one will hopefully lead to a better understanding of the practical considerations for optimising the algorithm in general settings.

For fixed  $\tau$  and  $A$ , we evaluate the performance of the reversible jump algorithm using mean absolute deviations (MADs) around quantities that are usually of interest for users: the posterior mode of  $K$  (denoted by  $k^*$ ), and the posterior mean and standard deviation of  $X_i^K$ ,  $i \in \{1, \dots, n + K\}$  (we consider the first parameter  $X_1^K$  for the simulation study). For a given sample produced by the reversible jump algorithm,  $k^*$  is estimated by  $\widehat{k}^*$ , the mode of the sample related to the random variable  $K$ , and  $\mu$  and  $\sigma$  are estimated by  $\widehat{\mu}$  and  $\widehat{\sigma}$ , which are respectively the mean and standard deviation of the sample related to the random variable  $X_i^K$ . Representative samples lead to accurate estimates, thus resulting in small absolute deviations. For fixed  $\tau$  and  $A$ , we approximate the MADs by  $\sum_{i=1}^N |\widehat{k}_i^* - k^*|/N$ ,  $\sum_{i=1}^N |\widehat{\mu}_i - \mu|/N$  and  $\sum_{i=1}^N |\widehat{\sigma}_i - \sigma|/N$ , where  $N$  is the number of samples, and  $\widehat{k}_i^*$ ,  $\widehat{\mu}_i$  and  $\widehat{\sigma}_i$  are respectively the mode, mean and standard deviation based on the sample  $i$ . We also compute a global measure that mimics the one used in Section 3 (and given in (8)) to optimise the algorithm with respect to  $\tau$ . This global measure is a linear combination of a standardised version of the MADs:

$$\frac{\frac{1}{N} \sum_{i=1}^N |\widehat{k}_i^* - k^*|}{\sqrt{\frac{1}{N-1} \sum_{i=1}^N (\widehat{k}_i^* - k^*)^2}} + \frac{1}{2} \left( \frac{\frac{1}{N} \sum_{i=1}^N |\widehat{\mu}_i - \mu|}{\sqrt{\frac{1}{N-1} \sum_{i=1}^N (\widehat{\mu}_i - \mu)^2}} + \frac{\frac{1}{N} \sum_{i=1}^N |\widehat{\sigma}_i - \sigma|}{\sqrt{\frac{1}{N-1} \sum_{i=1}^N (\widehat{\sigma}_i - \sigma)^2}} \right).$$

In this simulation study,  $N = 1,000$ ,  $\mu = 0$ ,  $\sigma = 1$ , and each sample is of size 100,000. The results are presented in Figure 3.

As expected, the performance of the algorithm regarding the estimation of the posterior distribution of  $K$  decreases as the value of  $\tau$  increases due to fewer model switchings. An increase in  $\tau$  has the opposite effect regarding the estimation of the posterior of  $\mathbf{X}^K$ . The vertical lines represent the optimal values for  $\tau$ , which are 0.415, 0.334 and 0.194, when  $A = 2, 5, 25$ , respectively. These values are optimal in the sense that they allow to attain the appropriate balance between an efficient estimation of the posterior of  $K$  (but a poor estimation of the posterior of  $\mathbf{X}^K$ ) and an efficient estimation of the posterior of  $\mathbf{X}^K$  (but a poor estimation of the posterior of  $K$ ).

Figure 3 also helps illustrate that reversible jump users should favor the smallest admissible value for  $A$ , an aspect that has been theoretically justified in Section 3. Indeed, the value of  $A$  has a direct impact on the performance regarding the estimation of the posterior distribution of  $K$  (the performance decreases as the value of  $A$  increases), and it has an indirect impact on the performance regarding the estimation of the posterior of  $\mathbf{X}^K$  through the optimal value for  $\tau$ .

We finally note that, as  $n \rightarrow \infty$ , the curves defined by the global measure as a function of  $\tau$  look like those in Figure 2 (a). However, for moderate values of  $n$ , the curves defined by the global measure are almost flat between 0.2 and 0.6. It indicates that, for moderate values of  $n$ , selecting any value in this range for  $\tau$  is almost optimal. As  $n$  increases, users should narrow down to the optimal value of  $\tau$ , especially for large value of  $A$ .

## 5. Conclusion

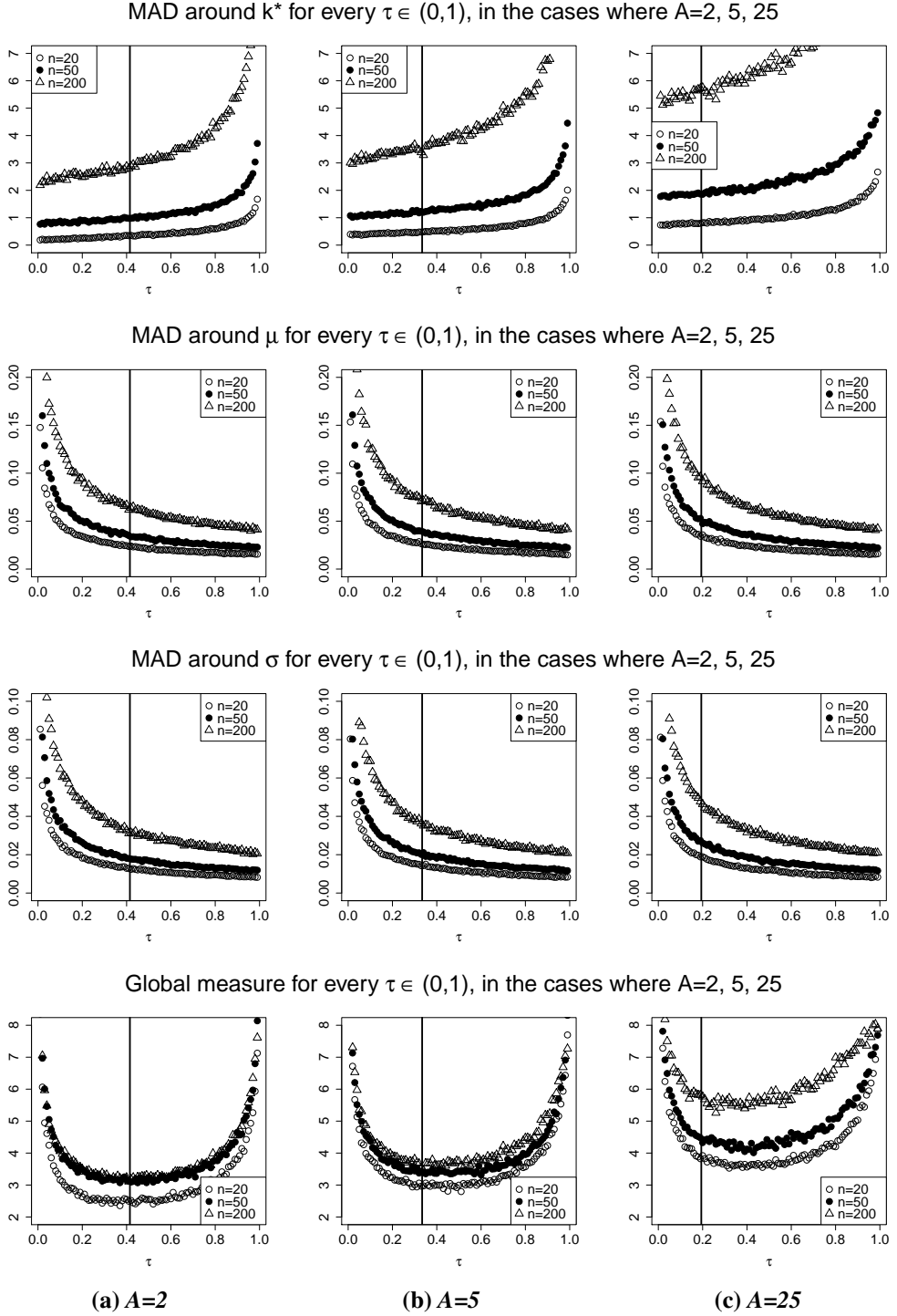
In this paper, we have provided guidelines to practically and efficiently implement the reversible jump algorithm described in Section 2. The performance of this algorithm depends on the inputs required for its implementation: the constants  $\ell$ ,  $\tau$  and  $A$ , and the proposal distribution  $q$ . The theoretical results derived in Section 3 allow to optimally choose values for  $\ell$  and  $\tau$ . The optimal value for  $\ell$  is given by  $2.38/\sqrt{\Upsilon}$ , which corresponds to an acceptance rate of candidates for updating the parameters of approximately 0.234 (considering only iterations in which there have been an attempt to update the parameters). The optimal value for  $\tau$  can be determined, for a given  $A$ , using (9) (or by looking at Figure 2 (b)). In Section 4.2, it has been explained that, for moderate values of  $n$ , selecting any value between 0.2 and 0.6 for  $\tau$  is almost optimal. The practical guidelines given in Section 4.1 enable to suitably design the proposal distribution  $q$ , which implicitly allows to determine the constant  $A$ .

The theoretical results hold when the algorithm is applied to sample from a target distribution  $\pi_n$  that satisfies the assumptions provided in Section 2. Essentially, the target distribution  $\pi_n$  must be a product of the PMF  $p$  (the distribution of the model indicator  $K$ ) and a  $(n + K)$ -product of PDFs  $f$  (the optimal value for  $\tau$  arising from (9) is theoretically valid when  $f = \mathcal{N}(\mu, \sigma^2)$ ). Being aware that this sampling context is simple, our goal was to make a first step towards automated implementation of the reversible jump algorithm. The distribution  $\pi_n$  is more often comprised of a product of different functions  $f_i$ , as in a context of selection of the principal components when the principal component regression is used to model the data. In Section 4.1, we have proposed an heuristic approach to optimally design the sampler when the algorithm is applied to sample from more elaborate target distributions.

## 6. Proof of Theorem 1

This section is dedicated to the demonstration of the main result of this paper, the weak convergence  $\{\mathbf{Z}_{1,2}^n(t), t \geq 0\} \Rightarrow \{\mathbf{Z}(t), t \geq 0\}$  in the Skorokhod topology as  $n \rightarrow \infty$  (the stochastic processes  $\{\mathbf{Z}_{1,2}^n(t), t \geq 0\}$  and  $\{\mathbf{Z}(t), t \geq 0\}$  have been defined in Theorem 1). Thus consider the sampling context described in Section 2.

In order to prove the result, we demonstrate the convergence of the finite-dimensional distributions of  $\{\mathbf{Z}_{1,2}^n(t), t \geq 0\}$  to those of  $\{\mathbf{Z}(t), t \geq 0\}$ . To achieve this, we verify condition (c) of Theorem 8.2 from Chapter 4 of Ethier and Kurtz (1986). The weak convergence then follows from Corollary 8.6 of Chapter 4 of Ethier and Kurtz (1986). The remaining conditions of Theorem 8.2 and the conditions specified in Corollary 8.6 are either straightforward or easily derived from the proof given in this section. They are nevertheless explicitly verified in Gagnon (2016).



**Fig. 3.** MADs around  $k^*$ ,  $\mu$  and  $\sigma$ , and the global measure, for every  $\tau \in (0,1)$ , in the cases where  $A = 2, 5, 25$  (the vertical lines represent the optimal values for  $\tau$ , which are 0.415, 0.334 and 0.194, when  $A = 2, 5, 25$ , respectively)

The proof of the convergence of the finite-dimensional distributions relies on the convergence of (what we call) the “pseudo-generator”, a quantity that we now introduce. The proof follows in Section 6.2.

### 6.1. Pseudo-Generator

In this section, we introduce a quantity that we call the “pseudo-generator” of  $\{\mathbf{Z}_{1,2}^n(t), t \geq 0\}$  due to its similarity with the infinitesimal generator of stochastic processes. It is defined as follows:

$$\varphi_n(t) := n\mathbb{E}[h(\mathbf{Z}_{1,2}^n(t + 1/n)) - h(\mathbf{Z}_{1,2}^n(t)) \mid \mathcal{F}^{\mathbf{Z}^n}(t)],$$

where  $h \in \mathcal{C}_c^\infty(\mathbb{R}^2)$ , the space of infinitely differentiable functions on  $\mathbb{R}^2$  with compact support. Theorem 2.5 from Chapter 8 of [Ethier and Kurtz \(1986\)](#) allows us to restrict our attention to this set of functions when studying the limiting behaviour of the pseudo-generator (see [Gagnon \(2016\)](#) for more details).

Let

$$R^K(\lfloor nt \rfloor) := \frac{K(\lfloor nt \rfloor) - \lfloor \sqrt{n} \log n \rfloor / 2}{\sqrt{n}} = Z_1^n(t).$$

The pseudo-generator  $\varphi_n(t)$  can be decomposed into three parts, each associated with a specific type of movement, as follows:

$$\varphi_n(t) = \varphi_{1,n}(t) + \varphi_{2,n}(t) + \varphi_{3,n}(t),$$

where  $\varphi_{1,n}(t)$  is associated with the update of the parameters, i.e.

$$\varphi_{1,n}(t) := n\tau\mathbb{E}\left[\left(h(R^K, Y_1^K) - h(R^K, X_1^K)\right) \left(1 \wedge \frac{\prod_{i=1}^{n+K} f(Y_i^K)}{\prod_{i=1}^{n+K} f(X_i^K)}\right) \mid R^K, \mathbf{X}^K\right],$$

$\varphi_{2,n}(t)$  is associated with the inclusion of an extra parameter, i.e.

$$\begin{aligned} \varphi_{2,n}(t) := & \frac{n(1-\tau)A}{A+1}\mathbb{E}\left[\left(h(R^{K+1}, X_1^K) - h(R^K, X_1^K)\right) \right. \\ & \times \left. \left(1 \wedge \frac{f(U)p(K+1)}{q(U)p(K)A}\right) \mid R^K, \mathbf{X}^K\right], \end{aligned} \quad (10)$$

and  $\varphi_{3,n}(t)$  is associated with the withdrawal of the last parameter, i.e.

$$\begin{aligned} \varphi_{3,n}(t) := & \frac{n(1-\tau)}{A+1}\mathbb{E}\left[\left(h(R^{K-1}, X_1^K) - h(R^K, X_1^K)\right) \right. \\ & \times \left. \left(1 \wedge \frac{q(X_{n+K}^K)p(K-1)A}{f(X_{n+K}^K)p(K)}\right) \mid R^K, \mathbf{X}^K\right]. \end{aligned} \quad (11)$$

Note that the Markov process  $\{(R^K(m), \mathbf{X}^{K(m)}(m)), m \in \mathbb{N}\}$  is time-homogeneous, and consequently, the time index has been omitted to simplify the notation. Also note that, when there is an update of the parameters, only the parameters  $\mathbf{X}^K$  move (the model indicator remains the same). When an extra parameter is included or the last parameter withdrawn, only the model indicator changes, as a switch from model  $K$  to model  $K+1$  or  $K-1$  is made.

## 6.2. Proof of the Convergence of the Finite-Dimensional Distributions

Condition (c) of Theorem 8.2 essentially reduces to the following convergence:

$$\mathbb{E} [|\varphi_n(t) - Gh(\mathbf{Z}_{1,2}^n(t))|] \rightarrow 0 \text{ as } n \rightarrow \infty,$$

where  $G$  is the generator of a diffusion with  $G = G_1 + G_2$  and

$$\begin{aligned} G_2 h(\mathbf{Z}_{1,2}^n(t)) &= \frac{1-\tau}{A+1} \times -Z_1^n(t) h_x(\mathbf{Z}_{1,2}^n(t)) + \frac{1-\tau}{A+1} h_{xx}(\mathbf{Z}_{1,2}^n(t)), \\ G_1 h(\mathbf{Z}_{1,2}^n(t)) &= \tau \ell^2 \Phi \left( -\frac{\ell \sqrt{\Upsilon}}{2} \right) (\log f(Z_2^n(t)))' h_y(\mathbf{Z}_{1,2}^n(t)) + \tau \ell^2 \Phi \left( -\frac{\ell \sqrt{\Upsilon}}{2} \right) h_{yy}(\mathbf{Z}_{1,2}^n(t)). \end{aligned}$$

The function  $h$  above is the same function  $h$  involved in the definition of the random variable  $\varphi_n(t)$  (given in Section 6.1). In other words, the convergence has to be proved for an arbitrary function  $h \in \mathcal{C}_c^\infty(\mathbb{R}^2)$ . The functions  $h_x$  and  $h_{xx}$  respectively represent the first and second derivatives of  $h$  with respect to its first argument. Analogously, the functions  $h_y$  and  $h_{yy}$  respectively represent the first and second derivatives of  $h$  with respect to its second argument. Note that it exists a positive constant  $M$  such that  $h$  and all its derivatives are bounded in absolute value by this constant.

Using the triangle inequality, we have

$$\begin{aligned} \mathbb{E} [|\varphi_n(t) - Gh(\mathbf{Z}_{1,2}^n(t))|] &\leq \mathbb{E} [|\varphi_{1,n}(t) - G_1 h(\mathbf{Z}_{1,2}^n(t))|] \\ &\quad + \mathbb{E} [|\varphi_{2,n}(t) + \varphi_{3,n}(t) - G_2 h(\mathbf{Z}_{1,2}^n(t))|]. \end{aligned}$$

In this paper, we show that the second term on the right-hand side (RHS) converges towards 0 as  $n \rightarrow \infty$ . The proof that the first term converges towards 0 is similar to that of Theorem 1.1 of [Roberts et al. \(1997\)](#), and is thus omitted (it can nevertheless be found in [Gagnon \(2016\)](#)).

The key here is the use of Taylor expansions in order to obtain derivatives of  $h$  as in generators of diffusions.

We first analyse  $\varphi_{2,n}(t)$  as defined in (10). As explained in Section 2,  $0 \leq p(K+1)/p(K) \leq 2$ , and therefore,

$$\frac{f(U)p(K+1)}{q(U)p(K)A} \leq \frac{2f(U)}{q(U)A} \leq 1.$$

Consequently, since  $h(R^{K+1}, X_1^K) = h(R^K + 1/\sqrt{n}, X_1^K)$ ,

$$\begin{aligned} \varphi_{2,n}(t) &= \frac{n(1-\tau)}{A+1} \left( h(R^K + 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) \right) \frac{p(K+1)}{p(K)} \mathbb{E} \left[ \frac{f(U)}{q(U)} \mid R^K, \mathbf{X}^K \right] \\ &= \frac{n(1-\tau)}{A+1} \left( h(R^K + 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) \right) \frac{p(K+1)}{p(K)}. \end{aligned}$$

In the last equality, we use the fact that  $U$  is independent of  $(K, \mathbf{X}^K)$ , and therefore,

$$\mathbb{E} \left[ \frac{f(U)}{q(U)} \mid R^K, \mathbf{X}^K \right] = \mathbb{E} \left[ \frac{f(U)}{q(U)} \right] = \int_{-\infty}^{\infty} \frac{f(u)}{q(u)} q(u) du = 1.$$

Note that  $\varphi_{2,n}(t) = 0$  when  $K = \lfloor \sqrt{n} \log n \rfloor$  since  $p(\lfloor \sqrt{n} \log n \rfloor + 1) = 0$ .

We now study  $\varphi_{3,n}(t)$  as defined in (11). As explained in Section 2, when  $2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor$



we have that  $p(K-1)/p(K) \geq 1/2$ . Therefore, when  $2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor$ ,

$$\frac{q(X_{n+K}^K)p(K-1)A}{f(X_{n+K}^K)p(K)} \geq \frac{q(X_{n+K}^K)A}{2f(X_{n+K}^K)} \geq 1.$$

This means that the acceptance probability of withdrawing the last parameter is 1, when it is possible to withdraw a parameter. Consequently, since  $h(R^{K-1}, X_1^K) = h(R^K - 1/\sqrt{n}, X_1^K)$ ,

$$\varphi_{3,n}(t) = \frac{n(1-\tau)}{A+1} \left( h(R^K - 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) \right) \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor).$$

Note that  $\varphi_{3,n}(t) = 0$  when  $K = 1$  since  $p(0) = 0$ .

By using Taylor expansions of  $h$  around  $R^K$ , we obtain that

$$\begin{aligned} h(R^K + 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) &= \frac{1}{\sqrt{n}} h_x(R^K, X_1^K) + \frac{1}{2n} h_{xx}(R^K, X_1^K) \\ &\quad + \frac{1}{6n^{3/2}} h_{xxx}(W, X_1^K), \end{aligned}$$

$$\begin{aligned} h(R^K - 1/\sqrt{n}, X_1^K) - h(R^K, X_1^K) &= -\frac{1}{\sqrt{n}} h_x(R^K, X_1^K) + \frac{1}{2n} h_{xx}(R^K, X_1^K) \\ &\quad - \frac{1}{6n^{3/2}} h_{xxx}(T, X_1^K), \end{aligned}$$

where  $W$  belongs to  $(R^K, R^K + 1/\sqrt{n})$ ,  $T$  belongs to  $(R^K - 1/\sqrt{n}, R^K)$ , and  $h_{xxx}$  represents the third derivative of  $h$  with respect to its first argument.

Therefore,

$$\begin{aligned} \varphi_{2,n}(t) + \varphi_{3,n}(t) - G_2 h(\mathbf{Z}_{1,2}^n(t)) &= \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \\ &\quad \times \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \left[ \sqrt{n} \left( \frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \\ &\quad + \mathbb{1}(K=1) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \left( \sqrt{n} \times \frac{p(K+1)}{p(K)} + R^K \right) \\ &\quad - \mathbb{1}(K = \lfloor \sqrt{n} \log n \rfloor) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) (\sqrt{n} - R^K) \\ &\quad + \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \frac{1-\tau}{2(A+1)} h_{xx}(R^K, X_1^K) \left( \frac{p(K+1)}{p(K)} - 1 \right) \\ &\quad + \mathbb{1}(K=1) \frac{1-\tau}{2(A+1)} h_{xx}(R^K, X_1^K) \left( \frac{p(K+1)}{p(K)} - 2 \right) \\ &\quad - \mathbb{1}(K = \lfloor \sqrt{n} \log n \rfloor) \frac{1-\tau}{2(A+1)} h_{xx}(R^K, X_1^K) \\ &\quad + \frac{1-\tau}{6\sqrt{n}(A+1)} h_{xxx}(W, X_1^K) \frac{p(K+1)}{p(K)} \mathbb{1}(1 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \\ &\quad - \frac{1-\tau}{6\sqrt{n}(A+1)} h_{xxx}(T, X_1^K) \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor). \end{aligned} \tag{12}$$

We now show that the expectation of the absolute value of each term on the RHS in (12) converges towards 0 as  $n \rightarrow \infty$ . Consequently, using the triangle inequality we will obtain

$$\mathbb{E} \left[ \left| \varphi_{2,n}(t) + \varphi_{3,n}(t) - G_2 h(\mathbf{Z}_{1,2}^n(t)) \right| \right] \rightarrow 0 \text{ as } n \rightarrow \infty.$$

We start with the last terms in (12) and make our way up. It is clear that the expectation of the absolute value of each of the last two terms converges towards 0 as  $n \rightarrow \infty$  since  $|h_{xxx}| \leq M$  and  $0 \leq p(K+1)/p(K) \leq 2$ .

We now analyse the fourth one (starting from the bottom). As  $n \rightarrow \infty$ ,

$$\begin{aligned} \mathbb{E} \left[ \left| \mathbb{1}(K=1) \frac{1-\tau}{2(A+1)} h_{xx}(R^K, X_1^K) \left( \frac{p(K+1)}{p(K)} - 2 \right) \right| \right] \\ \leq \frac{M(1-\tau)}{A+1} \times \mathbb{E} [\mathbb{1}(K=1)] = \frac{M(1-\tau)}{A+1} \times \mathbb{P}(K=1) \rightarrow 0, \end{aligned}$$

using  $|h_{xx}| \leq M$  and  $0 \leq |p(K+1)/p(K) - 2| \leq 2$  in the first inequality. Proposition 3 in Section 7 is then used to conclude that  $\mathbb{P}(K=1) \rightarrow 0$  as  $n \rightarrow \infty$ . The proof for the third term (starting from the bottom) is similar.

Applying Lemmas 1 to 3 from Section 7, each of the remaining terms is seen to converge towards 0 in  $\mathcal{L}^1$  as  $n \rightarrow \infty$ , and thus

$$\mathbb{E} \left[ \left| \varphi_{2,n}(t) + \varphi_{3,n}(t) - G_2 h(\mathbf{Z}_{1,2}^n(t)) \right| \right] \rightarrow 0 \text{ as } n \rightarrow \infty.$$

## 7. Results Used in the Proof of Theorem 1

**Lemma 1.** *As  $n \rightarrow \infty$ , we have*

$$\mathbb{E} \left[ \left| \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \frac{1-\tau}{2(A+1)} h_{xx}(R^K, X_1^K) \left( \frac{p(K+1)}{p(K)} - 1 \right) \right| \right] \rightarrow 0.$$

*Proof of lemma 1.* First,

$$\begin{aligned} \mathbb{E} \left[ \left| \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \frac{1-\tau}{2(A+1)} h_{xx}(R^K, X_1^K) \left( \frac{p(K+1)}{p(K)} - 1 \right) \right| \right] \\ \leq \frac{M(1-\tau)}{2(A+1)} \mathbb{E} \left[ \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \left| \frac{p(K+1)}{p(K)} - 1 \right| \right], \end{aligned}$$

because  $|h_{xx}| \leq M$ .

Considering the case where  $\lfloor \sqrt{n} \log n \rfloor$  is odd, we have

$$\begin{aligned} \mathbb{E} \left[ \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \left| \frac{p(K+1)}{p(K)} - 1 \right| \right] \\ = \mathbb{E} \left[ \mathbb{1} \left( \frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \left| \frac{p(K+1)}{p(K)} - 1 \right| \right] \\ + \mathbb{E} \left[ \mathbb{1} \left( 2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \left| \frac{p(K+1)}{p(K)} - 1 \right| \right]. \end{aligned}$$

We analyse each term separately. The first one corresponds to the case where  $K$  is at the mode or at

its right. Therefore,  $p(K+1)/p(K) = a_{K,n}$  and

$$\begin{aligned}
& \mathbb{E} \left[ \mathbf{1} \left( \frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \left| \frac{p(K+1)}{p(K)} - 1 \right| \right] \\
&= \mathbb{E} \left[ \mathbf{1} \left( \frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) |a_{K,n} - 1| \right] \\
&= \mathbb{E} \left[ \mathbf{1} \left( \frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \frac{|K - \lfloor \sqrt{n} \log n \rfloor / 2|}{n} \right] \\
&\leq \frac{\lfloor \sqrt{n} \log n \rfloor / 2 - 1}{n} \leq \frac{\log n}{2\sqrt{n}} \rightarrow 0,
\end{aligned}$$

because  $a_{K,n} = 1 - b_{K,n}/\sqrt{n}$  and  $b_{K,n} = |K - \lfloor \sqrt{n} \log n \rfloor / 2| / \sqrt{n}$ . We now study the case where  $K$  is at the left of the mode. Therefore,  $p(K+1)/p(K) = a_{K,n}^{-1}$  and

$$\begin{aligned}
& \mathbb{E} \left[ \mathbf{1} \left( 2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \left| \frac{p(K+1)}{p(K)} - 1 \right| \right] \\
&= \mathbb{E} \left[ \mathbf{1} \left( 2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) |a_{K,n}^{-1} - 1| \right] \\
&= \mathbb{E} \left[ \mathbf{1} \left( 2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \left| \frac{b_{K,n}}{\sqrt{n} - b_{K,n}} \right| \right] \leq \frac{(\log n)/2}{\sqrt{n} - (\log n)/2} \rightarrow 0,
\end{aligned}$$

using similar mathematical arguments as above and the fact that  $1/(2\sqrt{n}) \leq b_{K,n} \leq (\log n)/2$  when  $2 \leq K \leq (\lfloor \sqrt{n} \log n \rfloor - 1)/2$ . The proof for the case where  $\lfloor \sqrt{n} \log n \rfloor$  is even is similar. ■

**Lemma 2.** As  $n \rightarrow \infty$ , we have

$$\mathbb{E} \left[ \mathbf{1}(K=1) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \left( \sqrt{n} \times \frac{p(K+1)}{p(K)} + R^K \right) \right] \rightarrow 0,$$

and

$$\mathbb{E} \left[ \mathbf{1}(K = \lfloor \sqrt{n} \log n \rfloor) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) (\sqrt{n} - R^K) \right] \rightarrow 0.$$

*Proof of Lemma 2.* Using Proposition 3,  $|h_x| \leq M$ ,  $0 \leq p(K+1)/p(K) \leq 2$  and  $R^K := (K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n}$ , we have

$$\begin{aligned}
& \mathbb{E} \left[ \mathbf{1}(K=1) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \left( \sqrt{n} \times \frac{p(K+1)}{p(K)} + R^K \right) \right] \\
&\leq \frac{(1-\tau)M}{A+1} \left( 2\sqrt{n} + \frac{\log n}{2} \right) \mathbb{P}(K=1) \rightarrow 0,
\end{aligned}$$

since

$$\begin{aligned}
\mathbf{1}(K=1) \left| \sqrt{n} \times \frac{p(K+1)}{p(K)} + R^K \right| &\leq \mathbf{1}(K=1) \left( 2\sqrt{n} + \left| \frac{1 - \lfloor \sqrt{n} \log n \rfloor / 2}{\sqrt{n}} \right| \right) \\
&\leq \mathbf{1}(K=1) \left( 2\sqrt{n} + \frac{\log n}{2} \right).
\end{aligned}$$

The proof that

$$\mathbb{E} \left[ \left| \mathbb{1}(K = \lfloor \sqrt{n} \log n \rfloor) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) (\sqrt{n} - R^K) \right| \right] \rightarrow 0$$

is similar. ■

**Lemma 3.** *As  $n \rightarrow \infty$ , we have*

$$\mathbb{E} \left[ \left| \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \left[ \sqrt{n} \left( \frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \right| \right] \rightarrow 0.$$

*Proof of Lemma 3.* First,

$$\begin{aligned} & \mathbb{E} \left[ \left| \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \frac{1-\tau}{A+1} h_x(R^K, X_1^K) \left[ \sqrt{n} \left( \frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \right| \right] \\ & \leq \frac{(1-\tau)M}{A+1} \mathbb{E} \left[ \left| \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \left[ \sqrt{n} \left( \frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \right| \right], \end{aligned}$$

because  $|h_x| \leq M$ . Considering the case where  $\lfloor \sqrt{n} \log n \rfloor$  is odd, we have

$$\begin{aligned} & \mathbb{E} \left[ \left| \mathbb{1}(2 \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1) \left[ \sqrt{n} \left( \frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \right| \right] \\ & = \mathbb{E} \left[ \left| \mathbb{1} \left( \frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \left[ \sqrt{n} \left( \frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \right| \right] \\ & \quad + \mathbb{E} \left[ \left| \mathbb{1} \left( 2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \left[ \sqrt{n} \left( \frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \right| \right]. \end{aligned}$$

We analyse each term separately. The first one corresponds to the case where  $K$  is at the mode or at its right. Therefore,  $p(K+1)/p(K) = a_{K,n}$  and

$$\begin{aligned} & \mathbb{E} \left[ \left| \mathbb{1} \left( \frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \left[ \sqrt{n} \left( \frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \right| \right] \\ & = \mathbb{E} \left[ \left| \mathbb{1} \left( \frac{\lfloor \sqrt{n} \log n \rfloor + 1}{2} \leq K \leq \lfloor \sqrt{n} \log n \rfloor - 1 \right) \left[ \sqrt{n} (a_{K,n} - 1) + R^K \right] \right| \right] \\ & = \mathbb{E} \left[ \left| \mathbb{1} \left( \frac{1}{2\sqrt{n}} \leq R^K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 2}{2\sqrt{n}} \right) \left[ -b_{K,n} + R^K \right] \right| \right] = 0, \end{aligned}$$

because  $a_{K,n} = 1 - b_{K,n}/\sqrt{n}$  and  $b_{K,n} = |K - \lfloor \sqrt{n} \log n \rfloor/2|/\sqrt{n} = R^K$  when  $R^K \geq 0$  ( $R^K := (K - \lfloor \sqrt{n} \log n \rfloor/2)/\sqrt{n}$ ). We now study the case where  $K$  is at the left of the mode. Therefore,  $p(K+1)/p(K) = a_{K,n}^{-1}$  and

$$\begin{aligned} & \mathbb{E} \left[ \left| \mathbb{1} \left( 2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \left[ \sqrt{n} \left( \frac{p(K+1)}{p(K)} - 1 \right) + R^K \right] \right| \right] \\ & = \mathbb{E} \left[ \left| \mathbb{1} \left( 2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \left[ \sqrt{n} (a_{K,n}^{-1} - 1) + R^K \right] \right| \right] \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E} \left[ \mathbf{1} \left( 2 \leq K \leq \frac{\lfloor \sqrt{n} \log n \rfloor - 1}{2} \right) \left| \frac{\sqrt{n}}{\sqrt{n} - b_{K,n}} \times b_{K,n} + R^K \right| \right] \\
&= \mathbb{E} \left[ \mathbf{1} \left( \frac{4 - \lfloor \sqrt{n} \log n \rfloor}{2\sqrt{n}} \leq R^K \leq \frac{-1}{2\sqrt{n}} \right) \times -R^K \left| \frac{\sqrt{n}}{\sqrt{n} - b_{K,n}} - 1 \right| \right] \\
&\leq \mathbb{E} \left[ \mathbf{1} \left( \frac{4 - \lfloor \sqrt{n} \log n \rfloor}{2\sqrt{n}} \leq R^K \leq \frac{-1}{2\sqrt{n}} \right) \frac{\log n}{2} \times \frac{b_{K,n}}{\sqrt{n} - b_{K,n}} \right] \\
&\leq \frac{\log n}{2} \times \frac{(\log n)/2}{\sqrt{n} - (\log n)/2} \rightarrow 0,
\end{aligned}$$

using similar mathematical arguments as above, the fact that  $b_{K,n} = -R^K$  when  $R^K < 0$ , and that  $1/(2\sqrt{n}) \leq -R^K \leq (\log n)/2$  when  $(4 - \lfloor \sqrt{n} \log n \rfloor)/(2\sqrt{n}) \leq R^K \leq -1/(2\sqrt{n})$ . The proof for the case where  $\lfloor \sqrt{n} \log n \rfloor$  is even is similar. ■

**Proposition 3.** *The random variable  $K$  with PMF  $p$  defined in Section 2 is such that, for all  $\rho \in \mathbb{R}$ ,*

$$\lim_{n \rightarrow \infty} n^\rho \mathbb{P}(K = 1) = \lim_{n \rightarrow \infty} n^\rho \mathbb{P}(K = \lfloor \sqrt{n} \log n \rfloor) = 0.$$

*Proof of Proposition 3.* Consider the case where  $\lfloor \sqrt{n} \log n \rfloor$  is even. Using equation (4), we have

$$p(1) = p(\lfloor \sqrt{n} \log n \rfloor) = p\left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2}\right) \prod_{i=1}^{\frac{\lfloor \sqrt{n} \log n \rfloor}{2} - 1} \left(1 - \frac{i}{n}\right).$$

In the proof of Proposition 1 in Section 8, we show that  $p(\lfloor \sqrt{n} \log n \rfloor/2) \rightarrow 1/\sqrt{2\pi}$  as  $n \rightarrow \infty$ . Also, using the fact that  $1 - x \leq \exp\{-x\}$  for all  $x \in \mathbb{R}$ , we have for all  $\rho \in \mathbb{R}$

$$\begin{aligned}
n^\rho \prod_{i=1}^{\frac{\lfloor \sqrt{n} \log n \rfloor}{2} - 1} \left(1 - \frac{i}{n}\right) &\leq n^\rho \exp \left\{ - \sum_{i=1}^{\frac{\lfloor \sqrt{n} \log n \rfloor}{2} - 1} \frac{i}{n} \right\} \\
&= n^\rho \exp \left\{ - \frac{1}{2} \left( \frac{\frac{\lfloor \sqrt{n} \log n \rfloor}{2} - 1}{\sqrt{n}} \right)^2 - \frac{\frac{\lfloor \sqrt{n} \log n \rfloor}{2} - 1}{2n} \right\} \\
&\leq n^\rho \exp \left\{ - \frac{1}{2} \left( \frac{\lfloor \sqrt{n} \log n \rfloor}{2\sqrt{n}} - 1 \right)^2 \right\} \rightarrow 0,
\end{aligned}$$

as  $n \rightarrow \infty$ . Similarly, we can show the result for the case where  $\lfloor \sqrt{n} \log n \rfloor$  is odd. ■

## 8. Proofs of Propositions 1 and 2

*Proof of Proposition 1.* The random variable  $Z_1^n(t)$  is defined as  $(K(\lfloor nt \rfloor) - \lfloor \sqrt{n} \log n \rfloor/2)/\sqrt{n}$ , and  $K(\lfloor nt \rfloor) \sim p$  for all  $t$  and for all  $n$  (see Section 2 for the assumptions on  $p$ ). Therefore, to simplify the notation, the time index is omitted for the rest of the proof. Consider the constant  $z < 0$

and the case where  $\lfloor \sqrt{n} \log n \rfloor$  is even. We have

$$\begin{aligned} \mathbb{P}((K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n} \leq z) &= \mathbb{P}(K \leq z\sqrt{n} + \lfloor \sqrt{n} \log n \rfloor / 2) \\ &= \sum_{k=\lceil -z\sqrt{n} \rceil}^{\lfloor \sqrt{n} \log n \rfloor / 2 - 1} p(\lfloor \sqrt{n} \log n \rfloor / 2 - k) \\ &= p\left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2}\right) \sum_{k=\lceil -z\sqrt{n} \rceil}^{\lfloor \sqrt{n} \log n \rfloor / 2 - 1} \prod_{i=1}^k \left(1 - \frac{i}{n}\right), \end{aligned}$$

using  $\lfloor z\sqrt{n} + \lfloor \sqrt{n} \log n \rfloor / 2 \rfloor = \lfloor z\sqrt{n} \rfloor + \lfloor \sqrt{n} \log n \rfloor / 2 = \lfloor \sqrt{n} \log n \rfloor / 2 - \lceil -z\sqrt{n} \rceil$  in the second equality ( $\lceil \cdot \rceil$  is the ceiling function), and equation (4) in the last equality. The sum above is well-defined if  $n \geq \exp(-2z + 5)$  and we select  $n$  large enough to ensure this. Using again equation (4) and the fact that  $\sum_{k=1}^{\lfloor \sqrt{n} \log n \rfloor} p(k) = 1$ , we have

$$p\left(\frac{\lfloor \sqrt{n} \log n \rfloor}{2}\right) = \left(2 \left(1 + \sum_{k=1}^{\lfloor \sqrt{n} \log n \rfloor / 2 - 1} \prod_{i=1}^k \left(1 - \frac{i}{n}\right)\right)\right)^{-1}.$$

Therefore,

$$\mathbb{P}\left(\frac{K - \lfloor \sqrt{n} \log n \rfloor / 2}{\sqrt{n}} \leq z\right) = \frac{(1/\sqrt{n}) \sum_{k=\lceil -z\sqrt{n} \rceil}^{\lfloor \sqrt{n} \log n \rfloor / 2 - 1} \prod_{i=1}^k (1 - i/n)}{\frac{2}{\sqrt{n}} + \frac{2}{\sqrt{n}} \sum_{k=1}^{\lfloor \sqrt{n} \log n \rfloor / 2 - 1} \prod_{i=1}^k (1 - i/n)}.$$

Using the fact that  $1 - x \leq \exp\{-x\}$  for all  $x \in \mathbb{R}$ , we have

$$\prod_{i=1}^k \left(1 - \frac{i}{n}\right) \leq \exp\left\{-\sum_{i=1}^k \frac{i}{n}\right\} = \exp\left\{-\frac{1}{2} \left(\frac{k}{\sqrt{n}}\right)^2 - \frac{k}{2n}\right\}.$$

In addition, for all  $\delta > 0$ , there exists  $\epsilon > 0$  such that  $\exp\{-(1 + \delta)x\} \leq 1 - x$  for  $0 \leq x < \epsilon$ . Therefore, since  $0 \leq i/n \leq \lfloor \sqrt{n} \log n \rfloor / (2n) - 1/n \leq \log n / (2\sqrt{n}) \rightarrow 0$  as  $n \rightarrow \infty$  when  $1 \leq i \leq k \leq \lfloor \sqrt{n} \log n \rfloor / 2 - 1$ , for all  $\delta > 0$ , there exists a constant  $N > 0$  such that for all  $n \geq N$ ,

$$\prod_{i=1}^k \left(1 - \frac{i}{n}\right) \geq \exp\left\{-\frac{(1 + \delta)}{2} \left(\frac{k}{\sqrt{n}}\right)^2 - \frac{k(1 + \delta)}{2n}\right\}.$$

The objective is to use a ‘‘Riemann sum’’ argument, where the length of the subintervals of the partition is  $1/\sqrt{n}$ , to study the asymptotic behaviour of the numerator and denominator of  $\mathbb{P}((K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n} \leq z)$ . More precisely, we now prove that the numerator of  $\mathbb{P}((K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n} \leq z)$  converges towards  $\int_{-z}^{\infty} \exp(-x^2/2) dx$  and that the denominator converges towards  $\int_{-\infty}^{\infty} \exp(-x^2/2) dx = \sqrt{2\pi}$ . To achieve this, we use Lebesgue’s dominated convergence theorem.

First, we rewrite the numerator as

$$\frac{1}{\sqrt{n}} \sum_{k=\lceil -z\sqrt{n} \rceil}^{\lfloor \frac{\sqrt{n} \log n}{2} \rfloor - 1} \prod_{i=1}^k \left(1 - \frac{i}{n}\right) = \int_{-z}^{\infty} \sum_{k=\lceil -z\sqrt{n} \rceil}^{\lfloor \frac{\sqrt{n} \log n}{2} \rfloor - 1} \prod_{i=1}^k \left(1 - \frac{i}{n}\right) \mathbb{1}_{\left[\frac{k}{\sqrt{n}}, \frac{k+1}{\sqrt{n}}\right)}(x) dx.$$

Now, we analyse the integrand. For all  $x \in (-z, \infty)$  and for large enough  $n$ , there exists a unique  $k' \in \{\lceil -z\sqrt{n} \rceil, \dots, \lfloor \sqrt{n} \log n \rfloor / 2 - 1\}$  with  $\mathbb{1}_{[k'/\sqrt{n}, (k'+1)/\sqrt{n})}(x) = 1$ . Also,  $0 \leq x - k'/\sqrt{n} < 1/\sqrt{n}$ , which implies that  $k'/\sqrt{n} \rightarrow x$  as  $n \rightarrow \infty$ . Consequently, using the upper bound and the lower bound on  $\prod_{i=1}^k (1 - i/n)$ ,

$$\sum_{k=\lceil -z\sqrt{n} \rceil}^{\lfloor \frac{\sqrt{n} \log n}{2} \rfloor - 1} \prod_{i=1}^k \left(1 - \frac{i}{n}\right) \mathbb{1}_{\left[\frac{k}{\sqrt{n}}, \frac{k+1}{\sqrt{n}}\right)}(x) = \prod_{i=1}^{k'} \left(1 - \frac{i}{n}\right) \rightarrow \exp\{-x^2/2\},$$

as  $n \rightarrow \infty$ , because  $k'/n \leq \lfloor \sqrt{n} \log n \rfloor / (2n) - 1/n \leq \log n / (2\sqrt{n}) \rightarrow 0$ . Now, we prove that the integrand is bounded by an integrable function that does not depend on  $n$ . For all  $x \in (-z, \infty)$ ,

$$\begin{aligned} \prod_{i=1}^k \left(1 - \frac{i}{n}\right) \mathbb{1}_{\left[\frac{k}{\sqrt{n}}, \frac{k+1}{\sqrt{n}}\right)}(x) &\leq \exp\left\{-\frac{1}{2} \left(\frac{k}{\sqrt{n}}\right)^2 - \frac{k}{2n}\right\} \mathbb{1}_{\left[\frac{k}{\sqrt{n}}, \frac{k+1}{\sqrt{n}}\right)}(x) \\ &\leq \exp\left\{-\frac{1}{2}(x-1)^2\right\} \mathbb{1}_{\left[\frac{k}{\sqrt{n}}, \frac{k+1}{\sqrt{n}}\right)}(x), \end{aligned}$$

using the upper bound on  $\prod_{i=1}^k (1 - i/n)$  in the first inequality, and then  $x \leq (k+1)/\sqrt{n} \leq (k/\sqrt{n}) + 1$ . As a result,

$$\begin{aligned} \sum_{k=\lceil -z\sqrt{n} \rceil}^{\lfloor \frac{\sqrt{n} \log n}{2} \rfloor - 1} \prod_{i=1}^k \left(1 - \frac{i}{n}\right) \mathbb{1}_{\left[\frac{k}{\sqrt{n}}, \frac{k+1}{\sqrt{n}}\right)}(x) &\leq \exp\left\{-\frac{1}{2}(x-1)^2\right\} \sum_{k=\lceil -z\sqrt{n} \rceil}^{\lfloor \frac{\sqrt{n} \log n}{2} \rfloor - 1} \mathbb{1}_{\left[\frac{k}{\sqrt{n}}, \frac{k+1}{\sqrt{n}}\right)}(x) \\ &= \exp\left\{-\frac{1}{2}(x-1)^2\right\} \mathbb{1}_{\left[\frac{\lceil -z\sqrt{n} \rceil}{\sqrt{n}}, \frac{\lfloor \sqrt{n} \log n \rfloor / 2}{\sqrt{n}}\right)}(x) \\ &\leq \exp\left\{-\frac{1}{2}(x-1)^2\right\}, \end{aligned}$$

which is integrable. Similarly, we can prove that the denominator of  $\mathbb{P}((K - \lfloor \sqrt{n} \log n \rfloor / 2) / \sqrt{n} \leq z)$  converges towards  $\int_{-\infty}^{\infty} \exp(-x^2/2) dx = \sqrt{2\pi}$ , and we can show the result for  $z \geq 0$  and for the case where  $\lfloor \sqrt{n} \log n \rfloor$  is odd.  $\blacksquare$

*Proof of Proposition 2.* The random variables  $K(m)$  and  $U(m+1)$  are independent and such that  $K(m) \sim p$  and  $U(m+1) \sim q$  for all  $m \in \mathbb{N}$  (see Section 2 for the assumptions on  $p$  and  $q$ ). Therefore, to simplify the notation, the time index is omitted for the rest of the proof. As explained in Section 6.2,

$$\mathbb{E} \left[ 1 \wedge \frac{f(U)p(K+1)}{q(U)p(K)A} \right] = \mathbb{E} \left[ \frac{f(U)p(K+1)}{q(U)p(K)A} \right],$$

and  $\mathbb{E}[f(U)/q(U)] = 1$ . Finally, using Proposition 3, we have

$$\frac{1}{A} \mathbb{E} \left[ \frac{p(K+1)}{p(K)} \right] = \frac{1}{A} \sum_{k=1}^{\lfloor \sqrt{n} \log n \rfloor - 1} p(k+1) = \frac{1}{A} (1 - p(1)) \rightarrow \frac{1}{A} \text{ as } n \rightarrow \infty. \quad \blacksquare$$

## References

- Al-Awadhi, F., Hurn, M. and Jennison, C. (2004) Improving the acceptance rate of reversible jump mcmc proposals. *Statist. Probab. Lett.*, **69**, 189–198.
- Bédard, M. (2007) Weak convergence of metropolis algorithms for non-i.i.d. target distributions. *Ann. Appl. Probab.*, **17**, 1222–1244.
- Bédard, M. (2008) Optimal acceptance rates for metropolis algorithms: Moving beyond 0.234. *Stochastic Process. Appl.*, **118**, 2198–2222.
- Bédard, M., Douc, R. and Moulines, E. (2012) Scaling analysis of multiple-try mcmc methods. *Stochastic Process. Appl.*, **122**, 758–786.
- Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, Jesus-Maria and Stuart, A. (2013) Optimal tuning of the hybrid monte carlo algorithm. *Bernoulli*, **19**, 1501–1534.
- Beskos, A., Roberts, G. and Stuart, A. (2009) Optimal scalings for local metropolis–hastings chains on nonproduct targets in high dimensions. *Ann. Appl. Probab.*, **19**, 863–898.
- Bibby, B. M., Skovgaard, I. M. and Sørensen, M. (2005) Diffusion-type models with given marginal distribution and autocorrelation function. *Bernoulli*, **11**, 191–220.
- Brooks, S. P., Giudici, P. and Roberts, G. O. (2003) Efficient construction of reversible jump markov chain monte carlo proposal distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **65**, 3–39.
- Ethier, S. N. and Kurtz, T. G. (1986) *Markov Processes: Characterization and Convergence*, vol. 282. Wiley.
- Gagnon, P. (2016) *Robust Model Selection: Linear Regression and Reversible Jump Algorithm*. Ph.D. thesis, Université de Montréal (in preparation).
- Green, P. J. (1995) Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, **82**, 711–732.
- Hastie, D. (2005) *Towards Automatic Reversible Jump Markov Chain Monte Carlo*. Ph.D. thesis, University of Bristol.
- Hastings, W. K. (1970) Monte carlo sampling methods using markov chains and their applications. *Biometrika*, **57**, 97–109.
- Kang, B. (2013) Fast determinantal point process sampling with application to clustering. In *Advances in Neural Information Processing Systems*, 2319–2327.
- Karagiannis, G. and Andrieu, C. (2013) Annealed importance sampling reversible jump mcmc algorithms. *J. Comp. Graph. Stat.*, **22**, 623–648.
- Mattingly, J. C., Pillai, N. S. and Stuart, A. (2012) Diffusion limits of the random walk metropolis algorithm in high dimensions. *Ann. Appl. Probab.*, **22**, 881–930.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087.
- Neal, P. and Roberts, G. (2006) Optimal scaling for partially updating mcmc algorithms. *Ann. Appl. Probab.*, **16**, 475–515.
- Peskun, P. (1973) Optimum monte-carlo sampling using markov chains. *Biometrika*, **60**, 607–612.
- Richardson, S. and Green, P. J. (1997) On bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **59**, 731–792.
- Robert, C. and Casella, G. (2004) *Monte Carlo Statistical Methods*. Springer Science & Business Media.
- Roberts, G. O., Gelman, A. and Gilks, W. R. (1997) Weak convergence and optimal scaling of random walk metropolis algorithms. *Ann. Appl. Probab.*, **7**, 110–120.
- Roberts, G. O. and Rosenthal, J. S. (2001) Optimal scaling for various metropolis-hastings algorithms. *Statist. Sci.*, **16**, 351–367.
- Tierney, L. (1998) A note on metropolis-hastings kernels for general state spaces. *Ann. Appl. Probab.*, **8**, 1–9.